

Sampling with Walsh Transforms

Yi LU

Institute of Software,
Chinese Academy of Sciences,
Beijing, P.R. China
luyi666@gmail.com

Abstract

With the advent of massive data outputs at a regular rate, admittedly, signal processing technology plays an increasingly key role. Nowadays, signals are not merely restricted to physical sources, they have been extended to digital sources as well.

Under the general assumption of discrete statistical signal sources, we propose a practical problem of sampling incomplete noisy signals for which we do not know *a priori* and the sample size is bounded. We approach this sampling problem by Shannon's channel coding theorem. We use an extremal binary channel with high probability of transmission error, which is rare in communication theory. Our main result demonstrates that it is the large Walsh coefficient(s) that characterize(s) discrete statistical signals, regardless of the signal sources. Note that this is a known fact in specific application domains such as images. By the connection of Shannon's theorem, we establish the necessary and sufficient condition for our generic sampling problem for the first time. Finally, we discuss the cryptographic significance of sparse Walsh transform.

Keywords. Walsh transform, Shannon's channel coding theorem, channel capacity, extremal binary channel, generic sampling.

1 Introduction

With the advent of massive data outputs regularly, we are confronted by the challenge of big data processing and analysis. Admittedly, signal processing has become an increasingly key technology. An open question is the sampling problem with the signals, for which we assume that we do not know *a priori*. Due to reasons of practical consideration, sampling is affected by possibly strong noise and/or the limited measurement precision. Assuming that the

signal source is not restricted to a particular application domain, we are concerned with a practical and generic problem to sample these noisy signals.

Our motivation arises from the following problem in modern applied statistics. Assume the discrete statistical signals in a general setting as follows. The samples, generated by an arbitrary (possibly noise-corrupted) source F , are 2^n -valued for a fixed n . We assume that the noise source generates uniformly-distributed samples¹. Note that our assumption on a general setting of discrete statistical signals is described by the assumption that F is an arbitrary yet fixed (not necessarily deterministic) function. It is known to be a hypothesis testing problem to test presence of any real signals. Traditionally, F is a deterministic function with small or medium input size. It is computationally easy to collect the *complete and precise* distribution f of F . Based on relative entropy (or Kullback-Leibler distance), the conventional approach (aka. the classic distinguisher in statistical cryptanalysis [18, 19]) solves the sampling problem, given the distribution f *a priori*. Nevertheless, in reality, F might be a function that we do not have the complete description, or it might be a non-deterministic function, or it might just have large input size. Thus, it is infeasible to collect the complete and precise distribution f . This gives rise to the new generic statistical sampling problem with discrete incomplete noisy signals, using bounded samples.

In this work, we show that we can solve the generic sampling problem as reliable as possible without knowing *a priori*. We approach this problem by the novel use of Shannon's channel coding theorem, which establishes the achievability of channel capacity. This allows to obtain a simple robust solution with an arbitrarily small probability of error. Note that in the conventional approach (i.e., the classic distinguisher), the problem statement is slightly different and the solution is of a different form. Our work uses the binary channel. The channel is assumed to have extremely high probability of transmission error (and we call it the extremal binary channel), which is rare in communication theory [17]. In particular, for the Binary Symmetric Channel (BSC) with crossover probability $(1-d)/2$ and d is small (i.e., $|d| \ll 1$), the channel capacity is approximately $d^2/(2 \log 2)$. Further, we construct a non-symmetric binary channel with crossover probability $(1-d)/2$ and $1/2$ respectively (and d is small). We show that the channel capacity is approximately $d^2/(8 \log 2)$.

Our main contributions are as follows. First, we present the generic sampling theorem. We show that for this extremal non-symmetric binary

¹For the pure digital signal source F , which is our research subject throughout this work, this assumption is justified by the maximum entropy principle [6, P278].

channel, Shannon's channel coding theorem can solve the generic sampling problem under the general assumption of statistical signal sources (i.e., no further assumption is made about signal sources). Specifically, the *necessary and sufficient* condition is given *for the first time* to sample the incomplete noisy signals with bounded sample size for signal detection. It is interesting to observe that the classical signal processing tool of Walsh transform [2,9] is essential: regardless of the real signal sources, the large Walsh coefficient(s) characterize(s) discrete statistical signals. Put other way, when sampling incomplete noisy signals of the same source multiple times, one can expect to see *repeatedly* those large Walsh coefficient(s) of same magnitude(s) at the fixed frequency position(s). Note that this is known in specific application domains such as images, voices etc. Clearly, our result shows strong connection between Shannon's theorem and Walsh transform. Both are the key innovative technologies in digital signal processing.

Secondly, our generic sampling theorem is naturally linked to the new area of compressive sensing [7]. Compressive sensing is based on the ground of sparse representation of signals in the transform domain. This enables powerful sampling techniques (with respect to the complexity of time-domain components for access and the time cost) for the purpose of signal recovery. Specifically, sparse Fourier transform has been the main research subject in this area. Most recently, studies on sparse Walsh transform follow [14,15]. Our preliminary work finds that in the most general case, sparse Walsh transform is linked (see [12,13]) to the maximum likelihood decoding problem for linear codes, which is known to be NP-complete.

The rest of the paper is organized as follows. In Section 2, we give preliminaries on Walsh transforms. In Section 3, we review Shannon's channel coding theorem. In Section 4, we translate Shannon's theorem in the case of extremal binary channels to hypothesis testing problems. Based on the results, we present our main sampling theorem in Section 5; we also discuss the cryptographic significance. We give concluding remarks in Section 6.

2 Walsh Transforms in Statistics

Given a real-valued function $f : GF(2)^n \rightarrow \mathbb{R}$, which is defined on an n -tuple binary vector of input, the Walsh transform of f , denoted by \hat{f} , is another real-valued function defined as

$$\hat{f}(i) = \sum_{j \in GF(2)^n} (-1)^{\langle i, j \rangle} f(j), \quad (1)$$

for all $i \in GF(2)^n$, where $\langle i, j \rangle$ denotes the inner product between two n -tuple binary vectors i, j . For later convenience, we give an alternative definition below. Given an input array $x = (x_0, x_1, \dots, x_{2^n-1})$ of 2^n reals in the time domain, the Walsh transform $y = \hat{x} = (y_0, y_1, \dots, y_{2^n-1})$ of x is defined by

$$y_i = \sum_{j \in GF(2)^n} (-1)^{\langle i, j \rangle} x_j,$$

for any n -tuple binary vector i . We call x_i (resp. y_i) the time-domain component (resp. transform-domain coefficient) of the signal with size 2^n . For basic properties and references on Walsh transforms, we refer to [9, 11].

Let f be a probability distribution of an n -bit random variable $\mathcal{X} = (X_n, X_{n-1}, \dots, X_1)$, where each $X_i \in \{0, 1\}$. Then, $\hat{f}(m)$ is the *bias* of the Boolean variable $\langle m, \mathcal{X} \rangle$ for any fixed n -bit vector m , which is often called the output *pattern* or *mask*. Here, recall that a Boolean random variable \mathcal{A} has *bias* ϵ , which is defined by $\epsilon = E[(-1)^{\mathcal{A}}] = \Pr(\mathcal{A} = 0) - \Pr(\mathcal{A} = 1)$. Hence, if \mathcal{A} is uniformly distributed, \mathcal{A} has bias 0. Obviously, the pattern m should be nonzero.

Walsh transforms were used in statistics to find dependencies within a multi-variable data set. In the multi-variable tests, each X_i indicates the presence or absence (represented by ‘1’ or ‘0’) of a particular feature in a pattern recognition experiment. Fast Walsh Transform (FWT) is used to obtain all coefficients $\hat{f}(m)$ in one shot. By checking the Walsh coefficients one by one and identifying the large² ones, we are able to tell the dependencies among X_i ’s.

3 Review on Shannon’s Channel Coding Theorem

We briefly review Shannon’s famous channel coding theorem (cf. [6]). First, we recall basic definitions of Shannon entropy. The entropy $H(X)$ of a discrete random variable X with alphabet \mathcal{X} and probability mass function $p(x)$ is defined by

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x).$$

²We use the convention in signal processing to refer to the large transform-domain coefficient d as the one with a large absolute value throughout the paper.

The joint entropy $H(X_1, \dots, X_n)$ of a collection of discrete random variables (X_1, \dots, X_n) with a joint distribution $p(x_1, x_2, \dots, x_n)$ is defined by

$$H(X_1, \dots, X_n) = - \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log_2 p(x_1, x_2, \dots, x_n).$$

Define the conditional entropy $H(Y|X)$ of a random variable Y given another X as

$$H(Y|X) = \sum_x p(x) H(Y|X = x).$$

The mutual information $I(X; Y)$ between two random variables X, Y is equal to $H(Y) - H(Y|X)$, which always equals $H(X) - H(X|Y)$. A communication channel is a system in which the output Y depends probabilistically on its input X . It is characterized by a probability transition matrix that determines the conditional distribution of the output given the input.

Theorem 1 (Shannon's Channel Coding Theorem). *Given a channel, denote the input, output by X, Y respectively. We can send information at the maximum rate C bits per transmission with an arbitrarily low probability of error, where C is the channel capacity defined by*

$$C = \max_{p(x)} I(X; Y), \quad (2)$$

and the maximum is taken over all possible input distributions $p(x)$.

For the binary symmetric channel (BSC) with crossover probability³ p , C can be expressed by (cf. [6]):

$$C = 1 - H(p) \text{ bits/transmission.} \quad (3)$$

Herein, we refer to the BSC with crossover probability $p = (1 + d)/2$ and d is small (i.e., $|d| \ll 1$) as an extremal BSC. We can prove for the channel capacity for an extremal BSC (see Appendix for proof):

Corollary 1 (extremal BSC). *Given a BSC channel with crossover probability $p = (1 + d)/2$, if d is small (i.e., $|d| \ll 1$), then, $C \approx c_0 \cdot d^2$, where the constant $c_0 = 1/(2 \log 2)$.*

Therefore, for an extremal BSC, we can send one bit with an arbitrarily low probability of error with the minimum number of transmissions $1/C = (2 \log 2)/d^2$, i.e., $O(1/d^2)$. In next section, we will translate Corollary 1 to two useful statistical results. Interestingly, note that in communication theory, this extremal BSC is rare because of its low efficiency [17] and we typically deal with $|d| \gg 0$.

³that is, the input symbols are complemented with probability p

4 Statistical Translations of Shannon's Theorem

Let X_0, X_1 denote the Boolean random variable with bias $+d, -d$ respectively (and we restrict ourselves to $|d| \ll 1$). Denote the probability distribution of X_0, X_1 by D_0, D_1 respectively. Let $D \in \{D_0, D_1\}$. We are given a binary sequence of random bits with length N , and each bit is independent and identically distributed (i.i.d.) following the distribution D . As a consequence of Shannon's channel coding theorem, we now solve a hypothesis testing problem in statistics: answer the minimum N required to decide whether $D = D_0$ or $D = D_1$ with an arbitrarily low probability of error.

We translate this problem into a BSC channel coding problem as follows. The inputs are transmitted through a BSC with error probability $p = (1 - d)/2$. By Shannon's channel coding theorem, with a minimum number of $N = 1/C$ transmissions, we can reliably (i.e., with an arbitrarily low probability of error) determine whether the input is '0' or '1'. The former case implies that the received sequence corresponds to the distribution D_0 (i.e., a bit '1' occurs in the output sequence with probability p), while the latter case implies that the received sequence corresponds to the distribution D_1 (i.e., a bit '0' occurs in the output sequence with probability p). This solves the problem stated above. Using Corollary 1 with $p = (1 - d)/2$ (for $|d| \ll 1$), we have $N = (2 \log 2)/d^2$, i.e., $O(1/d^2)$. Thus, we have just shown that Shannon's Channel Coding Theorem can be translated to solve the following hypothesis testing problem:

Theorem 2. *Assume that the boolean random variable \mathcal{A}, \mathcal{B} has bias $+d, -d$ respectively and d is small. We are given a sequence of random samples, which are i.i.d. following the distribution of either \mathcal{A} or \mathcal{B} . We can tell the sample source with an arbitrarily low probability of error, using the minimum number N of samples $(2 \log 2)/d^2$, i.e., $O(1/d^2)$.*

Further, the following variant is more frequently encountered in hypothesis testing, in which we have to deal with a biased distribution and a uniform distribution altogether.

Theorem 3. *Assume that the boolean random variable \mathcal{A} has bias d and d is small. We are given a sequence of random samples, which are i.i.d. following the distribution of either \mathcal{A} or a uniform distribution. We can tell the sample source with an arbitrarily low probability of error, using the minimum number N of samples $(8 \log 2)/d^2$, i.e., $O(1/d^2)$.*

Proof. It is clear that the construction of using a BSC in the proof of Theorem 2 does not work here, as the biases (i.e., $d, 0$ respectively) of the two

sources are non-symmetric. Thus, we propose to use Shannon's channel coding theorem with a non-symmetric binary channel rather than a BSC.

Assume the channel with the following transition matrix

$$p(y|x) = \begin{pmatrix} 1-p_e & p_e \\ 1/2 & 1/2 \end{pmatrix},$$

where $p_e = (1-d)/2$ and d is small. The matrix entry in the x th row and the y th column denotes the conditional probability that y is received when x is sent. So, the input bit 0 is transmitted by this channel with error probability p_e (i.e., the received sequence has bias d if input symbols are 0) and the input bit 1 is transmitted with error probability $1/2$ (i.e., the received sequence has bias 0 if input symbols are 1).

To compute the channel capacity C (i.e., to find the maximum) defined in (2), no closed form solution exist in general. Nonlinear optimization algorithms [1,3] are known to find a numerical solution. Below, we propose a simple method to give a closed form estimate C for our extremal binary channel. As $I(X;Y) = H(Y) - H(Y|X)$, we first compute $H(Y)$ by

$$H(Y) = H\left(p_0(1-p_e) + (1-p_0) \times \frac{1}{2}\right), \quad (4)$$

where p_0 denote $p(x=0)$ for short. Next, we compute $H(Y|X)$ as follows,

$$\begin{aligned} H(Y|X) &= \sum_x p(x)H(Y|X=x) \\ &= p_0\left(H(p_e) - 1\right) + 1. \end{aligned} \quad (5)$$

Combining (4) and (5), we have

$$I(X;Y) = H\left(p_0 \times \frac{1}{2} - p_0 p_e + \frac{1}{2}\right) - p_0 H(p_e) + p_0 - 1.$$

As $p_e = (1-d)/2$, we have

$$I(X;Y) = H\left(\frac{1+p_0 d}{2}\right) - p_0\left(H\left(\frac{1-d}{2}\right) - 1\right) - 1.$$

We apply (14) (in Appendix)

$$I(X;Y) = -\frac{p_0^2 d^2}{2 \log 2} - p_0\left(H\left(\frac{1-d}{2}\right) - 1\right) + O(p_0^4 d^4), \quad (6)$$

for small d . Note that the last term $O(p_0^4 d^4)$ on the right side of (6) is ignorable. Thus, $I(X; Y)$ approaches the maximum when

$$p_0 = -\frac{H(\frac{1-d}{2}) - 1}{d^2/(\log 2)} \approx \frac{d^2/(2 \log 2)}{d^2/(\log 2)} = \frac{1}{2}.$$

Consequently, we estimate the channel capacity from (6) by

$$C \approx -\frac{1}{4}d^2/(2 \log 2) + \frac{1}{2}\left(1 - H\left(\frac{1-d}{2}\right)\right) \approx -d^2/(8 \log 2) + d^2/(4 \log 2),$$

which is $d^2/(8 \log 2)$. □

Remark 1. *In statistical cryptanalysis (cf. [18, 19]), Theorem 2 and Theorem 3 were known in slightly different contexts: the probability of error is a parameter and the sample number is known on the order of $1/d^2$. By asking for an arbitrarily low probability of error, we are able to give an alternative proof using channel capacity rather than relative entropy (or Kullback-Leibler distance). While the latter is used as the classical tool to solve hypothesis testing problems, here we show that hypothesis testing problems can be linked to channel capacity.*

5 Sampling Theorems with Incomplete Signals

In this section, we apply the hypothesis testing result (Theorem 3) to two sampling problems (the classical and generic versions). Without loss of generality, we assume the discrete statistical signals are not restricted to a particular application domain. Assume that (possibly noise-corrupted) signals are 2^n -valued and noises are uniformly distributed. For the signal detection problem (i.e., to test presence of real signal), we adopt the conventional approach of statistical hypothesis testing. Rather than using the direct signal detection method (as done in specific application domains), we propose to perform the test between the associated distribution and the uniform distribution.

We give the mathematical model on the signal F as follows. F is an arbitrary (and not necessarily deterministic) function. Let X be the n -bit output sample of F , assuming that the input is random and uniformly distributed. Denote the output distribution of X by f . Note that our assumption on a general setting of discrete statistical signals is described by the assumption that F is an arbitrary yet fixed function.

Firstly, the classical sampling problem (which can be interpreted as the classical distinguisher⁴) is formally stated as follows.

Theorem 4 (Classical Sampling Problem). *Assume that the largest Walsh coefficient of f is $d = \hat{f}(m_0)$ for a nonzero n -bit vector m_0 . We can detect F with an arbitrarily low probability of error, using minimum number $N = (8 \log 2)/d^2$ of samples of F , i.e., $O(1/d^2)$.*

The proof can be easily obtained by applying Theorem 3 and we omit it here. The classical sampling problem assumes that F together with its characteristics (i.e., the largest Walsh coefficient d) are known *a priori*. It aims at detecting signal with an arbitrarily low probability of error, using minimum samples.

Next, we will present our main sampling theorem, a more practical (and widely applicable) sampling theorem formally. Assuming that it is infeasible to know signal F *a priori*, we want to detect signals with an arbitrarily low probability of error and with bounded sample size. Note that the sampled signal is incomplete (possibly noisy) and the associated distribution is noisy (i.e., not precise). And we call this problem as generic sampling with incomplete noisy signals. In contrast to the classical distinguisher, this result can be interpreted as a generalized distinguisher⁵ in the context of statistical cryptanalysis. We give our first result with $n = 1$ below.

Theorem 5 (Generic Sampling Problem with $n = 1$). *Assume that the sample size of F is upper-bounded by N . Regardless of the input size of F , in order to detect F with an arbitrarily low probability of error, it is necessary and sufficient to have the following condition satisfied, i.e., f has a nontrivial Walsh coefficient d with $|d| \geq c/\sqrt{N}$, where the constant $c = \sqrt{8 \log 2}$.*

Proof. Note that the only nontrivial Walsh coefficient d for $n = 1$ is $\hat{f}(1)$, which is nothing but the bias of F . First, we will show by contradiction that this is a necessary condition. That is, if we can identify F with an arbitrarily low probability of error, then, we must have $|d| \geq c/\sqrt{N}$. Suppose $|d| < c/\sqrt{N}$ otherwise. Following the proof of Theorem 3, we know that the error probability is bounded away from zero as the consequence of Shannon's Channel Coding Theorem. This is contradictory. Thus, we have shown that the condition on d is a necessary condition. Next, we will show that it is also

⁴As mentioned in Remark 1, the problem statement of the classical distinguisher is slightly different; it often deals with a large d (using a slightly different N) rather than the largest d (cf. [19]).

⁵With $n = 1$, this appears as an informal result in cryptanalysis, which is used as a black-box analysis tool in several crypto-systems.

a sufficient condition. That is, if $|d| \geq c/\sqrt{N}$, then, we can identify F with an arbitrarily low probability of error. This follows directly from Theorem 4 with $n = 1$. We complete our proof. \square

Now, we make a generalized proposition for $n \geq 1$, which incorporates Theorem 5 as a special case:

Proposition 1 (Generic Sampling Problem with $n \geq 1$). *Assume that the sample size of F is upper-bounded by N . Regardless of the input size of F , in order to detect F with an arbitrarily low probability of error, it is necessary and sufficient to have the following condition satisfied, i.e., $\sum_{i \neq 0} (\hat{f}(i))^2 \geq (8n \log 2)/N$.*

We note that the sufficient condition can be proved based on results of classic distinguisher (i.e., Squared Euclidean Imbalance) which uses the notion of relative distance and states that $\sum_{i \neq 0} (\hat{f}(i))^2 \geq (4n \log 2)/N$ is required for high probability [19].

According to Theorem 5 and Proposition 1, note that a real signal F should have the following property in the form of ℓ_2 norm of the associated distribution given the sample size N :

$$\|\hat{f}\|_2^2 \geq 1 + 8n \log 2/N,$$

where the ℓ_2 norm of f is defined as

$$\|f\|_2 = \sqrt{\sum_{i \in GF(2)^n} f(i)^2}.$$

By duality of time-domain and transform-domain signals, we make another proposition following Proposition 1:

Proposition 2. *The discrete statistical signals can be characterized by large Walsh coefficients of the associated distribution.*

Proposition 2 implies that the most significant transform-domain signals are the largest coefficients in our generalized model. This is a known fact in application domains such as images, voices etc. Nonetheless, for those signals, Walsh transform is directly applied to the time-domain samples rather than the associated distribution of the collected samples in our model; in analogy to Proposition 2, it is known that those signals can be characterized by large Walsh coefficients as well.

5.1 Cryptographic Significance on Sparse Walsh Transforms

In symmetric cryptanalysis, Walsh transforms play an essential role (cf. [5, 11]), including bias computing.

Following the recent successful development of compressive sensing [7], it is shown that surprisingly, sparse Fourier transform significantly outperforms FFT (Fast Fourier Transforms). For the problem size N , k -sparse Fourier transforms ($k \ll N$) aims at faster computing k non-zero or large coefficients and $(N - k)$ zero or negligible small ones, in comparison to FFT. For instance, according to [20, Fig. 1], with $N = 2^{28}$, $k = 50$, theoretical estimate on the time complexity of FFT is $N \cdot \log_2 N \approx 7 \times 10^9$ units; for sparse Fourier transforms, the estimated theoretical complexity is 10^7 units, i.e., a great reduction factor of 700 is obtained.

Due to the similarity of Fourier transform and Walsh transform, most recently, research on sparse Walsh transform follows [14, 15]. As illustration, assume k non-zero coefficients and $(N - k)$ zero coefficients in a simplified model. With the same parameters ($N = 2^{28}$, $k = 50$) as above, for sparse Walsh transform, the conservative theoretical time complexity⁶ is around 38000 units. This time unit is not comparable to the one in the case of FWT, i.e., 7×10^9 units. Nonetheless, we estimate a rough reduction factor of 8000 by [16, Fig. 8]. Additionally, for $k = 2, 4, 12, 25$, sparse Walsh transform [16] has the estimated time of 1600, 3000, 8200, 16400 units respectively.

According to our discussions in this section, it is natural to link the first key challenge to the generic approach of sparse Walsh transforms. In [12, 13], finding the largest Walsh coefficient is linked to maximum likelihood decoding problem for linear codes, which is known to be NP-complete. Assume k large coefficients and $(N - k)$ zero or negligible small ones in a general setting. It seems other than FWT, no efficient algorithms exist to compute sparse Walsh transforms. In contrast, in the simplified k -sparse model, theoretical estimate for the time complexity corresponding to $k = 1, 2$ is $(\log_2 N)^2, 2(\log_2 N)^2$. That is, we have the complexity on the order of $(\log_2 N)^2$ (resp. $N \log_2 N$) in the simplified model (resp. the general model). And we are working on approximate signal recovery in presence of noise to gain more insights about the first challenge.

⁶The required time-domain components for access is around 6700 (see [16, Theorem 1]) rather than N for FWT.

6 Concluding Remarks

We model general discrete statistical signals as the output samples of an unknown arbitrary yet fixed function (which is the signal source). We translate Shannon's channel coding theorem in the extremal case of a binary channel to solve a hypothesis testing problem. Due to high probability of transmission error, this extremal binary channel is rare in communication theory. Nonetheless, the translated result allows to solve a generic sampling problem, for which we know nothing about the signal source *a priori* and we can only afford bounded sampling measurements. Our main results demonstrate that the classical signal processing tool of Walsh transform is essential: it is the large Walsh coefficient(s) that characterize(s) discrete statistical signals, regardless of the signal sources. By Shannon's theorem, we establish the *necessary and sufficient* condition for the generic sampling problem under the general assumption of statistical signal sources. It shows strong connection between Shannon's theorem and Walsh transform; both are the key innovative technologies in digital signal processing. Our results can also be seen as generalization of the classic distinguisher; the latter is based on relative distance and is the standard tool for statistical hypothesis testing problems. Finally, based on our preliminary work on sparse Walsh transforms in the context of compressive sensing, we discuss the cryptographic significance.

References

- [1] S. Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Trans. Inform. Theory, IT-18 14-20, 1972.
- [2] J. M. Blackledge. Digital Signal Processing – Mathematical and Computational Methods, Software Development and Applications. Horwood Publishing, England, Second Edition, 2006.
- [3] R. Blahut, Computation of channel capacity and rate distortion functions. IEEE Trans. Inform. Theory, IT-18: 460-473, 1972.
- [4] E. J. Candès, T. Tao, Near optimal signal recovery from random projections: universal encoding strategies?, IEEE Trans. Inform. Theory, vol. 52, No. 12, pp. 5406-5425, Dec. 2006.

- [5] F. Chabaud, S. Vaudenay, Links between differential and linear cryptanalysis, EUROCRYPT 1994, LNCS vol. 950, pp. 356-365, Springer-Verlag, 1995.
- [6] T. M. Cover, J. A. Thomas. Elements of Information Theory. John Wiley & Sons, New York, 1991.
- [7] D. L. Donoho, Compressed sensing, IEEE Trans. Inform. Theory, vol. 52, No. 4, pp. 1289-1306, Apr. 2006.
- [8] R. M. Gray, L. D. Davisson. An Introduction to Statistical Signal Processing. Cambridge University Press, 2004. <http://www-ee.stanford.edu/~gray/sp.pdf>.
- [9] K. J. Horadam. Hadamard Matrices and Their Applications. Princeton University Press, 2007.
- [10] A. Joux. Algorithmic Cryptanalysis. Chapman & Hall/CRC, Cryptography and Network Security, 2009.
- [11] Y. Lu, Y. Desmedt, Walsh transforms and cryptographic applications in bias computing, https://sites.google.com/site/yilusite/Lu-Desmedt_preprint2015revised_ver.pdf, to appear in Cryptography and Communications, Springer.
- [12] Y. Lu, S. Vaudenay, Faster correlation attack on Bluetooth keystream generator E0, CRYPTO 2004, LNCS vol. 3152, pp. 407-425, Springer-Verlag, 2004.
- [13] Y. Lu, S. Vaudenay, Cryptanalysis of an E0-like combiner with memory, Journal of Cryptology, vol. 21, pp. 430-457, Springer (2008)
- [14] K. Ramchandran, PULSE: Peeling-based Ultra-low Complexity Algorithms for Sparse Signal Estimation, <http://www.eecs.berkeley.edu/~kannanr/project.html>.
- [15] R. Scheibler, S. Haghighatshoar, M. Vetterli, SparseFHT project, <https://github.com/LCAV/SparseFHT>.
- [16] R. Scheibler, S. Haghighatshoar, M. Vetterli, A Fast Hadamard Transform for Signals with Sub-linear Sparsity, Fifty-first Annual Allerton Conference, pp. 1250-1257, IEEE, Oct. 2013.
- [17] M. A. Shokrollahi, personal communication.

- [18] S. Vaudenay, An experiment on DES - statistical cryptanalysis, Third ACM Conference on Computer Security, pp. 139-147, 1996.
- [19] S. Vaudenay. A Classical Introduction to Modern Cryptography - Applications for Communications Security. Springer, New York, 2006.
- [20] C. Wang, M. Araya-Polo, S. Chandrasekaran et al., Parallel Sparse FFT, ACM, 2013.

Appendix: Proof of Corollary 1

Let $p = (1 + d)/2$ and so $|d| \leq 1$. For $|d| < 1$, we will first show

$$H\left(\frac{1+d}{2}\right) = 1 - \left(\frac{d^2}{2} + \frac{d^4}{12} + \frac{d^6}{30} + \frac{d^8}{56} + \underbrace{\dots}_{O(d^{10})}\right) \times \frac{1}{\log 2}. \quad (7)$$

We have

$$-H\left(\frac{1+d}{2}\right) = \frac{1+d}{2} \log_2 \frac{1+d}{2} + \frac{1-d}{2} \log_2 \frac{1-d}{2} \quad (8)$$

$$= \frac{1}{\log 2} \left(\frac{1+d}{2} \log \frac{1+d}{2} + \frac{1-d}{2} \log \frac{1-d}{2} \right) \quad (9)$$

$$= \frac{1}{\log 2} \left(\frac{1+d}{2} \log(1+d) + \frac{1-d}{2} \log(1-d) - \log 2 \right) \quad (10)$$

$$= \frac{1}{\log 2} \left(\frac{1}{2} \log(1-d^2) + \frac{d}{2} \log \frac{1+d}{1-d} - \log 2 \right) \quad (11)$$

by definition of entropy. Using Taylor expansion series for $0 \leq d < 1$, we have

$$\log(1-d^2) = -\left(d^2 + \frac{d^4}{2} + \frac{d^6}{3} + \frac{d^8}{4} + \dots\right) \quad (12)$$

$$\log \frac{1+d}{1-d} = 2\left(d + \frac{d^3}{3} + \frac{d^5}{5} + \frac{d^7}{7} + \dots\right) \quad (13)$$

Putting (12) and (13) into (11), we have

$$\begin{aligned} -H\left(\frac{1+d}{2}\right) &= \frac{1}{\log 2} \left(-\frac{1}{2} \left(d^2 + \frac{d^4}{2} + \frac{d^6}{3} + \frac{d^8}{4} + \dots \right) + \right. \\ &\quad \left. \left(d^2 + \frac{d^4}{3} + \frac{d^6}{5} + \frac{d^8}{7} + \dots \right) - \log 2 \right) \\ &= \frac{1}{\log 2} \left(\frac{d^2}{2} + \frac{d^4}{12} + \frac{d^6}{30} + \frac{d^8}{56} + \dots \right) - 1, \end{aligned}$$

which leads to (7) for $0 \leq d < 1$. For $-1 < d \leq 0$, we use symmetry of entropy $H(\frac{1+d}{2}) = H(\frac{1-d}{2})$ and apply above result to justify the validity of (7) for $|d| < 1$.

Note that if $|d| \ll 1$, (7) reduces to

$$H\left(\frac{1+d}{2}\right) = 1 - d^2/(2 \log 2) + O(d^4). \quad (14)$$

So, we can calculate C in (3) by

$$C = 1 - H\left(\frac{1+d}{2}\right) = \left(d^2 + O(d^4)\right)/(2 \log 2) \approx \frac{d^2}{2 \log 2},$$

which completes our proof.